

R E P O R T R E S U M E S

ED 003 306

24

EFFECTIVENESS OF A PILOT INFORMATION SERVICE FOR EDUCATIONAL RESEARCH MATERIALS.

BY- TAGUE, JEAN

WESTERN RESERVE UNIV., CLEVELAND

REPORT NUMBER CRP-1743

PUB DATE

63

WESTERN RESERVE UNIV., CLEVELAND, SCH.OF LIB. SCI.

EDRS PRICE MF-\$0.18 HC-\$2.52 63P.

DESCRIPTORS- *INFORMATION RETRIEVAL, *EDUCATIONAL RESEARCH,
*INFORMATION STORAGE, SYSTEMS ANALYSIS, CLEVELAND, OHIO

✓

REQUIREMENTS OF A PILOT INFORMATION SERVICE FOR EDUCATIONAL RESEARCH WERE INVESTIGATED BY ISOLATING THE CHARACTERISTICS OF DOCUMENTS JUDGED RELEVANT BY QUESTIONERS (USERS). IN ADDITION, RETRIEVAL STRATEGIES WERE COMPARED ON THE BASIS OF THE RELEVANCE AND RECALL FACTORS FOR EACH STRATEGY. THE FOLLOWING CONCLUSIONS WERE REACHED--(1) IN PROGRAMING QUESTIONS, THE NUMBER OF WORDS REQUIRED IN COMBINATION IS OF GREATER SIGNIFICANCE THAN THE TYPE OF WORDS REQUIRED, AS FAR AS RETRIEVAL OR RELEVANT AS OPPOSED TO PERIPHERAL OR NONRELEVANT ANSWERS IS CONCERNED, (2) THE RELATED WORDS WHICH A PROGRAMER ADDS TO AN ORIGINAL STATEMENT OF A QUESTION ARE USEFUL AS INDEXING AND EVALUATING WORDS, ALTHOUGH THEIR INDEXING EFFECTIVENESS IS LOWER THAN THAT OF QUESTION WORDS, (3) A SEMANTIC CODE LEADS FROM THE QUESTION WORDS TO VERY FEW, ADDITIONAL INDEXING TERMS, (4) TITLES NEED NOT BE RECORDED ON TAPE AS SEPARATE ENTITIES, AND (5) THE INFORMATION IN TELEGRAPHIC ABSTRACTS SHOULD BE INCLUDED IN CONVENTIONAL ABSTRACTS. (TC)

ED003306

REFERENCE
COPY

COOPERATIVE
RESEARCH
PROGRAM

No. 1743

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education

This document has been reproduced exactly as received from the
person or organization originating it. Points of view or opinions
stated do not necessarily represent official Office of Education
position or policy.

EFFECTIVENESS OF A PILOT INFORMATION SERVICE FOR EDUCATIONAL RESEARCH MATERIALS

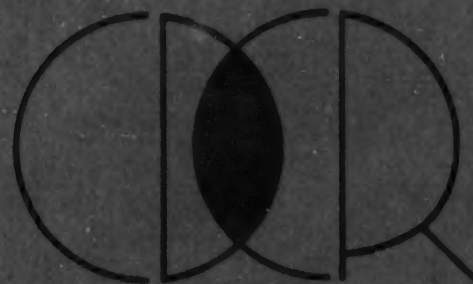
COOPERATIVE RESEARCH PROJECT NO. 1743

CENTER FOR DOCUMENTATION AND COMMUNICATION RESEARCH

SCHOOL OF LIBRARY SCIENCE

WESTERN RESERVE UNIVERSITY

CLEVELAND 6, OHIO



Approved
8-21-64
U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education

This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

**Effectiveness of a Pilot Information
Service for Educational
Research Materials**

Cooperative Research Project No. 1743

Jean Tague

**Center for Documentation and Communication Research
School of Library Science
Western Reserve University
Cleveland 6, Ohio**

1963

**The research reported herein was supported by the
Cooperative Research Program of the Office of Education,
U. S. Department of Health, Education and Welfare.**

Table of Contents

Preface	i
Introduction	1
Part I	4
--Purpose	4
--Procedure	5
--Condition of study	8
--Results	9
--Conclusions	31
Part II	34
--Purpose	35
--Procedure	35
--Results	36
--Conclusions	37
References	54

Tables

I. Total Answers Retrieved	10 a, b
II. Relevancy and Recall Factors	10 c
III. Causes for Nonretrieval of Relevant Answers	14 a
IV. Causes for Retrieval of Nonrelevant Answers	14 b
V. Effect of Role Indicators	21 a
VI. Results for All 14 Questions	36 a
VII. Results for 4 Questions Given 2 Outside Evaluations	36 b
VIII. Changes in Evaluation: All 14 Questions	36 c
IX. Changes in Evaluation: Outside Evaluation	36 d
X. Changes in Evaluation: 3 Evaluators	36 e
XI. Occurrence of Question Words, Semantic Code References and Program Additions	44 a
XII. Comparison of Question Words, Semantic Code References and Program Additions as Indexing Terms	47 a
XIII. Comparison of Question Words, Semantic Code References and Program Additions as Evaluating Terms	47 b
XIV. Results of Various Types of Program and Search Output	49 a
XV. Comparison of Title Words, Conventional Abstract Words and Telegraphic Abstract Words as Indexing Terms	50 a
XVI. Comparison of Title Words, Conventional Abstract Words and Telegraphic Abstract Words as Evaluating Terms	50 b

Appendixes

A. Alternate Strategies	55
-------------------------	----

Preface

There is a feeling of urgency in today's research environment: an urgency which results from public belief in the importance and the usefulness of research itself. That research is "useful" seems to mean that it should somehow be applied--the sooner the better--to a real-life situation. Applied research in science and technology will speed our journey to the moon. Applied research in information retrieval will produce the library of tomorrow--today.

But every element in an environment need not be a healthful one. We are told that the "air" on certain planets contains poison. And the element of urgency, in a research environment, can be breathed only sparingly and with care. Application, in this context, must be distinguished from exploitation.

To hedge against such unwise enthusiasm and premature claims, the Center for Documentation and Communication Research has intensified its interest in testing and test design. Directed toward several subject fields, this interest is occupying the time of most of the senior staff. Cleverdon's earlier, large-scale test, in the field of metallurgical literature, is referred to in Mrs. Tague's study. It was in many ways the "only begetter" of this report.

Mrs. Tague's careful work indicates both the limits (see pages 32 and 52) and the promise of early design. She has contributed to the formulation of the Center's thoughts on this important problem, and has aided in our planning for work with the Users Group which formed, in October, 1963, to serve in a dual role of guinea pig and (more properly) consultant in the months to come. Tests created and reported within the framework of a rigorous experimental design will illuminate the progress of many within the information field.

A. J. Goldwyn
Associate Director

**EFFECTIVENESS
OF A
PILOT INFORMATION SERVICE
FOR EDUCATIONAL RESEARCH MATERIALS**

Jean Tague

Introduction

During the year April, 1961, to April, 1962, the Center for Documentation and Communication Research, under contract to the U. S. Office of Education, established a pilot information service for educational research. This present study investigates the effectiveness of the service in satisfying the needs of a particular user group and suggests procedures for system improvement.

Two aspects of information retrieval were considered, for purposes of this study, to be of greatest significance in evaluating the pilot system:

- (1) determination of user needs
- (2) determination of the optimum method, within the present system, of satisfying these needs.

More specifically, user requirements were investigated by isolating the characteristics of documents judged relevant by the questioner, and retrieval strategies were compared on the basis of the relevance and recall factors for each strategy.

The sample used to investigate both aspects was obtained in the following manner. A set of 24 questions was selected from the more than 400 submitted during the initial year of the project by educators, researchers and other interested individuals, in response to a widely circulated

announcement of the pilot information service.¹ For comparative purposes, several strategies were used in programming each question (i. e., stating the question in the retrieval system language). Answers retrieved by these programs were sent to the original questioners and, in some cases, to others who had expressed interest in the same question. Each questioner was asked to evaluate the answers as relevant to his needs, peripheral or nonrelevant. These evaluations were used in comparing the alternate retrieval strategies and in investigating the characteristics of relevant as opposed to peripheral and nonrelevant documents. Details of the procedures and results of the alternate strategies are discussed in Part I of this report; details of the relevancy characteristics in Part II.

The two studies may be considered aspects of the same problem--what Becker and Hayes call the precision of the answer to a request:²

"The other aspect of the demands with respect to precision relates to the information recovered from the file and its degree of logical agreement with the requested information. Such agreement may be complete, fragmentary, uncertain; items may be missed or irrelevant to the requester. Each of these considerations--logical agreement, coverage, volume--must be specified in some way."

Part I of the report, from this point of view, is concerned with the recovery of information--i.e., relevancy and recall--and volume of answers; Part II with a rough quantification of the logical agreement between request and answer in relevant, peripheral and nonrelevant documents.

Pilot Operation

The unit operations of the educational research information service--acquisition, abstracting, indexing, coding, storage and retrieval--have been fully described in The Library of Tomorrow--Today.³ For convenience, we shall briefly review the main features.

The educational research file, at the beginning of this study, consisted of 4,837 documents published in the years 1932 to 1962. A "document" may be a periodical article, institutional or governmental research report, dissertation, book, or chapter of a book.

Two types of abstracts are written for each document. The conventional abstract is an English language summary of the article, giving bibliographic data and the significant contents of the article. It is designed to serve as the "product" of a search and a guide to the questioner in deciding whether or not he wishes to consult the original article. The telegraphic abstract analyzes the subject content of a document in a form amenable to machine searching. It consists of a list of the significant words of the article organized into syntactical relationships by "role indicators" and "level indicators," or "punctuation." The words in the telegraphic abstract are keypunched, matched against an English word-semantic code dictionary file and encoded automatically, unless the word is new, in which case a semantic code is assigned. The semantic code serves:

- (a) to make explicit the thesaural relationships of the terminology of the telegraphic abstract,
- (b) to convert the natural language of the telegraphic abstract into machine-readable form.

The encoded punched cards are used as input for the magnetic tape file of abstracts. Documents are also indexed by a faceted classification schedule developed during the pilot study from the Tauber-Lilley media classification system.⁴ The complete schedule is given in The Library of Tomorrow--Today.⁵ Both indexes may be searched by the computer for answers to an information request.

2. Semantic code programs were written for each question, incorporating index for the words of the question and words considered by the programmer to be related to the search. **PART I**

ALTERNATE SEARCH STRATEGIES by the semantic factors they contain, and then used as a thesaurus and partly from the programmers' **PURPOSE**

Our purpose in the effectiveness study was to compare the various codes were organized by topic indicators, levels, and the connections. approaches possible in retrieving answers from the pilot educational research file: narrow and broad semantic code programs and faceted classification. Questions programs may be represented by Boolean polynomials such as $(A + B) \cdot C$ or $(A + B) \cdot C + D$. $(A + B) \cdot C + D$ is indicated by the logic of the question. Comparisons are made on the basis of the relevancy and recall factors for each retrieval program, where, following Cleverdon,⁶ we define these factors, in relation to a program "A," as follows:

relevancy factor = $\frac{\text{number of relevant answers retrieved by Program A}}{\text{total number of answers retrieved by Program A}}$

recall factor = $\frac{\text{number of relevant answers retrieved by Program A}}{\text{total number of relevant answers retrieved by any program}}$

An example of the alternate strategies written for one question is given in Appendix A. These definitions differ from Cleverdon's in one respect: the recall factor is the ratio of relevant answers retrieved by a program to the total number of relevant answers retrieved by all programs. For Cleverdon, the recall factor is the ratio of relevant answers retrieved by a program to the total number of relevant answers retrieved by any program.

to total relevant answers in the file. This last figure, under our experimental conditions, could be obtained only by a complete scanning of the file of 4,837 documents by the original questioner.

Procedure

1. Twenty-four questions were selected from the more than 400 gathered during the initial phase of the pilot project.

2. Semantic code programs were written for each question, incorporating codes for the words of the question and words considered by the programmer to be related to the search. These additional terms were compiled partly from the reverse code dictionary (which groups words by the semantic factors they contain, and thus acts as a thesaurus) and partly from the programmers' intuition or general knowledge of the file. In these initial programs, codes were organized by role indicators, levels, and the connective "4" (or), "." (and), and "-" (but not), as indicated by the logic of the question. Question programs may be represented by Boolean polynomials such as
$$6 \left\{ 5 \left[(A + B) \cdot C \right]^5 + 4 \left[D \cdot (E + F) \right]^4 + G \right\}^6$$
 where the parentheses indicate levels of association and the letters stand for semantic factors (with or without infixes and end numbers), combinations of semantic factors, or role indicators. The initial semantic code programs are termed "N" programs, or narrow semantic code programs.

An example of the alternate strategies written for one question is given in Appendix A.

3. Broader semantic code programs, employing fewer of the discriminatory features of the system, were written for each question (B programs).

N programs were broadened by one or more of the following devices:

- a) eliminating role indicators.
- b) raising the level of required combinations.
- c) omitting one or more of the conjuncts of a conjunction--i.e., A.B rather than A.B.C, where the letters may represent a semantic factor or combination of semantic factors.
- d) adding alternate codes as disjuncts--i.e., A + B, rather than A alone.
- e) omitting infixes--i.e., S-CH instead of SUCH, where "-" is a universal character, allowing retrieval of this semantic factor with any infix.*

4. Programs using appropriate combinations from the faceted classification schedule were also written ("C" programs). This schedule allows documents to be classified by one or more of the following facets: population, media, teaching subjects and conditions, administration, and research design. Facets are coded by four-letter symbols and thus may be recorded on tape and machine searched in the same manner as the semantic coded telegraphic abstract. Faceted classification programs use the concepts rather than the words of the question and are generally quite broad, at least when compared with the narrow semantic code programs.

5. Narrow and broad semantic code programs were run against the entire test file of 4,837 documents. At the beginning of the study, faceted classifications had been assigned on tape only to a smaller set of 543 documents, and the C programs were therefore searched against this portion alone.

*Except the V infix, which must, for mechanical reasons, always be specified separately.

to apply the faceted classification codes to all documents in the file.

6. Answers were evaluated internally by staff members at the CDCR as relevant, peripheral, or nonrelevant. For this evaluation and the later evaluation by the questioner, these ratings were defined as follows:

relevant: an answer which has direct bearing on the question

peripheral: an answer of possible or partial bearing on the question

nonrelevant: an answer which has no bearing on the question

7. It was noticed that some of the relevant answers retrieved by the faceted classification program had been missed even by the broad semantic code program. For some questions, where these omissions appeared important, the B-programs were again revised so that the missed answers would be retrieved. These revised programs ("BR" programs) were also run against the total file.

8. Answers rated relevant or peripheral by the CDCR evaluator were sent to the original questioner for his evaluation. Ideally, of course, all answers, relevant, peripheral and nonrelevant, should have been sent to the questioner, but the large outputs for some of the B and BR programs made this procedure impractical.

9. Identical sets of answers with the same request for evaluation were sent to a number of other individuals who had expressed interest in similar subject areas. No contact was made with the questioner after the evaluations were received.

10. Evaluations were received from 14 of the 24 questioners. Second outside evaluations were received for four of these. Results given in this study are based on the 14 returned evaluations.

11. The questioners' evaluation verified our earlier conclusion that the C-programs were retrieving a number of relevant answers missed by the semantic code programs, both narrow and broad. It was therefore decided to apply the faceted classification codes to all documents in the file,

so that the classification would provide an alternate or supplementary method of searching. When all codes had been assigned and converted to tape, the entire file was searched with the C-programs and the results evaluated.

12. Relevancy and recall factors for N, B, BR and C programs were computed for each question, based both on the CDCR evaluation and the questioners' evaluation. The significance of the difference between programs, for both relevancy and recall, was tested by the Wilcoxon matched pairs test.

13. Based on the questioners' evaluation, relevant answers not retrieved by the original N-programs were examined to determine the causes of non-retrieval.

14. Also based on the questioners' evaluation, all nonrelevant answers retrieved by the original N-programs were examined to determine the causes of retrieval.

Conditions of Study

Before discussing the results of the alternate strategies, it might be well to stress the conditions of this test which distinguish it from most operational situations. No contact was made with the questioner after the initial statement of the question was received. In almost all cases, the brevity of the original statement left it open to a number of interpretations of increasing degrees of generality. Therefore, when answers were evaluated at the CDCR, those agreeing partially with the questioner's requirements were considered peripheral; only those answers with absolutely no relationship to the question were considered non-relevant. Of course, when answers were evaluated by the original questioner,

he was able to amplify his original specifications, and thus tended to be more selective in assigning the ratings "relevant" or "peripheral" to answers. These conditions may explain some of the discrepancies between evaluations by the CDCR and the questioner which are tabulated in Part II. Table II gives the relevancy and recall factors for each

type of program, based on the CDCR evaluation (Table IIA) and the questioner's evaluation (Table IIB). A dash (---) indicates relevancy or recall not retrieved by any program.

Results

The 14 questions which were evaluated and returned by the original questioners are listed below, with the numbers by which they will be identified throughout this report:

<u>Question #</u>	<u>Statement of question</u>
1	High school dropout as related to juvenile delinquency.
2	Pupil-directed learning (independent study or discovery).
3	Influence of parents on their children's school performance.
4	Decision-making.
5	Education of migratory children.
6	Role of teaching machines in treating mental illness.
7	Student interests, both stated and as measured by the interest inventory, and how they relate to achievement.
8	Use of audiovisual aids in teaching deaf children.
9	Meaning of achievement test results in terms of actual performance.
10	Programmed learning as compared with other teaching methods.
11	Teaching of film (film appreciation, literacy or comprehension).
12	Teaching of concepts by inductive or deductive techniques.
13	Effectiveness of learner participation in teaching motor skills by film.
14	Relationship of scholastic achievement to success on the job.

The total outputs for each program and the distribution of relevant, peripheral and nonrelevant documents according to the CDCR evaluation and the questioner evaluation are given in Table I: questions given one outside evaluation in Table IA and questions given two outside evaluations in Table IB. Table II gives the relevancy and recall factors for each type of program, based on the CDCR evaluation (Table IIA) and the questioners' evaluation (Table IIB). A dash (--) indicates relevancy or recall factors which are indeterminate--i.e., the relevancy factor for a program which retrieves no answers and the recall factor for a question for which no relevant answers were retrieved by any program.

The problem of variations in relevancy assessments by two or more evaluators will be considered in Part II. In the discussion which follows in this part only results from the questioners' evaluation will be used.

From Table IIB we see that the narrow semantic code program proved most efficient or equally efficient from the point of view of relevancy for 11 out of 18 questions (considering separate evaluations of the same question as different questions): questions 3 (a), 3 (b), 4 (a), 4 (b), 5 (a), 7, 8, 9, 10, 12 (a), 12 (b). For five of the remaining seven questions, the relevancy factor for the N-program is indeterminate, so that it cannot be compared with that of the other programs. In two cases--questions 11 and 14--no relevant answers were retrieved by any program. Of the remaining two questions, one, question 2, showed the highest relevancy factor with the C-program, and the other, question 5 (b), with the broad semantic code program.

It appears, then, that the narrow semantic code program is most efficient in producing a high proportion of relevant documents among those

Table I

Total Answers Retrieved, CDCR,

By Question Number and Program;

Evaluation by CDCR and Questioners

A. Questions Evaluated by CDCR and Original Questioner

	Question Number	Total Answers	Relevant Answers		Peripheral Answers		Nonrelevant Answers	
			CDCR	Questioner	CDCR	Questioner	CDCR	Questioner
(1) N-Program	1	0	0	0	0	0	0	0
(2) B-Program	2	52	34	14	7	17	11	21
	6	0	0	0	0	0	0	0
	7	29	11	11	8	7	10	11
	8	3	0	3	3	0	0	0
	9	41	18	7	18	23	5	11
(3) C-Program	10	327	296	244	0	24	31	59
	11	0	0	0	0	0	0	0
	13	0	0	0	0	0	0	0
	14	0	0	0	0	0	0	0
(2) B-Program	1	1	1	1	0	0	0	0
	2	75	36	24	24	25	15	26
	6	67	1	1	0	0	66	66
	7	32	13	11	8	8	11	13
	8	6	2	5	3	0	1	1
	9	363	22	11	32	37	309	315
	10	1242	322	260	5	30	915	952
	11	44	3	0	6	8	35	36
	13	13	1	3	8	4	4	6
	14	18	2	0	5	1	11	17
(3) BR-Program	1	16	3	4	1	0	12	12
	6	615	8	11	3	0	604	604
	11	209	3	0	50	41	156	168
	13	89	27	19	12	9	50	61
(4) C-Program	1	12	2	2	9	8	1	2
	2	29	15	11	11	14	3	4
	6	12	9	9	1	1	2	2
	7	68	11	10	17	18	40	40
	8	18	9	10	2	1	7	7
	9	59	18	10	20	27	21	22
	10	365	267	245	12	27	86	93
	11	61	0	0	17	14	44	47
	13	20	6	4	5	6	9	10
	14	33	0	0	9	9	24	24

Table I (cont.)

B. Questions Evaluated by CDCR,

Original Questioner (a) and Second Questioner (b)

	Question Number	Total Answers	Relevant Answers		Peripheral Answers		Nonrelevant Answers				
			CDCR	Questioner	CDCR	Questioner	CDCR	Questioner			
			(a)	(b)	(a)	(b)	(a)	(b)			
(1) N-Program	3	31	17	7	17	14	4	0	0	20	14
	4	115	22	5	11	50	21	33	43	89	71
	5	9	8	7	1	1	2	3	0	0	5
	12	81	30	12	7	37	22	2	14	47	72
(2) B-Program	3	283	54	16	59	69	24	16	160	243	208
	4	197	27	6	13	56	26	36	114	165	148
	5	25	13	9	2	2	6	4	10	10	19
	12	229	40	15	11	61	30	3	128	184	215
(3) C-Program	3	56	15	13	23	14	8	1	27	35	32
	4	84	3	0	2	14	12	13	67	72	69
	5	13	10	9	2	2	3	4	1	1	7
	12	77	18	9	4	14	11	5	45	57	68

B. Questioner Evaluation

(a) = Original Questioner; (b) = Second Questioner

1	0	1.00	.17	.33	.33
2	.27	.48	.32	.83	.38
3(a)	.23	.35	.96	.80	.23
3(b)	.55	.17	.21	.34	.37
4(a)	.84	.83	.43	1.00	.9
4(b)	.18	.45	.07	1.00	.15
5(a)	.78	.78	.36	1.00	1.00
5(b)	.11	.50	.34	1.00	1.00
6	0	.01	.07	.05	.15
7	.38	.68	.34	.63	.35
8	1.00	.25	.83	.42	.81
9	.17	.19	.03	.61	.37
10	.73	.90	.21	.96	.85
11	--	--	0	--	0
12(a)	.13	.80	.07	1.00	.65
12(b)	.09	.64	.05	1.00	.35
13	--	0	.23	.35	.19
14	--	--	0	--	--

Table II

Relevancy and Recall Factors

For Narrow, Broad and Broad Revised Semantic Code Programs

and Faceted Classification Programs

A. CDCR Evaluation

Question Number	N Program		B Program		BR Program		C Program	
	Relevancy	Recall	Relevancy	Recall	Relevancy	Recall	Relevancy	Recall
1	---	0	1.00	.20	.19	.60	.17	.40
2	.65	.81	.48	.88			.52	.37
3	.55	.29	.19	.93			.27	.26
4	.19	.81	.14	1.00			.04	.11
5	.89	.62	.52	1.00			.77	.78
6	---	0	.01	.11	.01	.89	.73	1.00
7	.38	.61	.41	.72			.16	.61
8	---	0	.33	.22			.50	1.00
9	.44	.60	.06	.73			.31	.60
10	.91	.89	.26	.97			.73	.80
11	---	0	.07	1.00	.01	1.00	0	0
12	.37	.75	.12	1.00			.23	.45
13	---	0	.08	.03	.30	.93	.30	.20
14	---	0	.11	1.00			0	0

B. Questioner Evaluation

(a) = Original Questioner; (b) = Second Questioner

1	questions 2, 3 (a), 0 (b), 4 (a), 1.00 4 (b), .17 5 (a), .23 7, 9, .67 11, .17	.33
2	.27 .48 .32 .83	.38
3(a)	.23 .35 .06 .80	.23
3(b)	.55 .27 .21 .94	.41
4(a)	.04 .83 .03 1.00	0
4(b)	.10 .85 .07 1.00	.02
5(a)	.78 .78 .36 1.00	.69
5(b)	.11 .50 .34 1.00	.13
6	these last four are 0 most specific of 16 questions, .02	.92
7	.38 .69 .34 .69	.15
8	machine 1.00 of two answers well defined c... .42	.56
9	.17 .39 .03 .61	.17
10	delinquency teaching machine .21 .96	.67
11	---	0
12(a)	.15 .80 .07 1.00	.12
12(b)	.09 .64 .03 1.00	.03
13	tion has ... phrase in very 23 .14	.21
14	---	0

which it retrieves. This result is as it should be, since the N-program is designed to be in closest logical agreement with the original question, using more discriminatory features and requiring a greater depth of indexing than the other programs. If one were guided only by relevancy considerations, a narrow semantic code program would always be preferable. However, considering the recall factor, the situation is almost completely reversed. Cleverdon⁷ has pointed out that a "weaker" (i.e., less discriminatory) system can match the recall of a more powerful system only at a lower relevance figure. The Center's experience has been that a less discriminatory system will have a higher recall figure, but, again, at the cost of a decrease in relevance. The broad revised semantic code programs, where these were written, had the highest recall factors, but this result was expected, since these programs were written, after evaluating the outputs from the C-program, to retrieve with the semantic code relevant answers which had been missed by the broad programs. If we disregard the BR-programs, we find that the broad semantic code programs had higher, or equal, recall factors for twelve of the eighteen questions: questions 2, 3 (a), 3 (b), 4 (a), 4 (b), 5 (a), 5 (b), 7, 9, 10, 11, 12 (a), 12 (b). For questions 11 and 14 the recall factors were indeterminate. For the remaining four questions, the faceted classification produced the highest recall rating: questions 1, 6, 8, 13. It is interesting that these last four are the most specific of the 14 questions, asking for a combination of two or more well-defined concepts: dropout and juvenile delinquency; teaching machines and mental illness; audiovisual aids and deaf children; learner participation, motor skills and film. If a question has been phrased in very specific terms, it appears that we can get

better results when the indexing vocabulary has been very strictly controlled, as in the faceted classification. In this connection, it is interesting to compare a statement by Becker and Hayes regarding the relative complexity of question and answer:⁸

Now it should be recognized that the complexity of the information recovered is not necessarily a reflection of the complexity with which the request was phrased. In fact, there is a very close relationship between these two aspects of complexity and the existence of a theory in the discipline of the stored information. In general, the better the theory the simpler the question which can be asked and the more complex the possible answer. If a very good theory exists, as in physics, then very simple questions can be asked and very complex answers received. On the other hand, with a poor theory one must ask very complex questions in order to recover even simple answers.

The mean relevance and recall factors for the four types of programs are as follows:

<u>Program</u>	<u>Mean Relevance</u>	<u>Mean Recall</u>
H	.35	.48
B	.23	.73
BR	.12	.83
C	.26	.54

The faceted classification appears to lie between the high relevance and low recall of the narrow semantic code programs, and the low relevance, high recall of the broad semantic code programs. However, it should be remembered that the H and B programs represent extreme cases of narrowness and broadness in using the semantic code, and that semantic code programs can be written at many levels of generality or specificity between these two.

At this point we should consider whether the differences between relevance and recall factors of H, B and C programs may be considered significant on the basis of a sample of 14 questions. Since at this time

we hesitate to assume a normal distribution for these factors, we shall use one of the nonparametric tests described by Siegel⁹--the Wilcoxon matched-pairs signed ranks test, applied to the difference scores between each set of relevancy and recall factors. The following results were obtained:

<u>Programs</u>	<u>Relevancy Factors</u>	<u>Recall Factors</u>
N and B programs	Significant difference at .05 level. No significant difference at .01 level.	Significant difference at .01 level--i.e., B has greater recall factor.
N and C programs	No significant difference at .05 level.	No significant difference at .05 level.
B and C programs	No significant difference at .05 level.	No significant difference at .05 level.

Apparently, a larger sample is needed.

Improving the Narrow Semantic Code Programs

The preceding results indicate that it would be preferable to use the narrow semantic code programs from the point of view of relevancy. However, their low recall factors--mean recall is only 48% of relevant documents--would probably not satisfy most questioners. Since the discriminatory features utilized by the N-programs already exist in the pilot file, it would seem most reasonable, as a first approach to system improvement, to attempt to raise the recall factor for the N-programs. Our first step, then, was to examine the relevant answers not retrieved by the N-programs, to determine how either the program or the abstract might have been altered so that these answers would have been retrieved. This examination was carried out for all missed relevant answers, and from it, some general causes of nonretrieval were detected.

Table III

Causes for Nonretrieval of Relevant Answers

The same procedure was followed in determining the causes of retrieval of nonrelevant answers by the narrow semantic code programs. All nonrelevant answers were examined and generalizations made, based on similarities among nonrelevant answers.

The purpose of these analyses was to discover the modifications which might enable the present telegraphic abstract-semantic code system to achieve its optimum relevancy and recall factors. Tables III and IV summarize the results for missed relevant answers and retrieved nonrelevant answers, respectively. The factors causing retrieval or nonretrieval, represented by the column headings in these tables, are discussed below, with examples, in order of decreasing frequency. Relevancy and nonrelevancy is based on the questioner's evaluation. In cases of two outside evaluations, relevant answers are those considered relevant by either evaluator; nonrelevant answers are those considered nonrelevant by both evaluators. These distinctions are made because the system operator wants to retrieve those answers considered relevant by any questioner and does not want to retrieve answers considered nonrelevant by all questioners.

Factors Causing Nonretrieval of Relevant Answers by N-Programs1. Terminology

Relevant answers were missed by the N-program because words in the telegraphic abstract relating to the inquiry were not predicted by the programmer from the question words, semantic code relationships, or intuitive knowledge of the file.

The semantic code, represented by three-letter "semantic factors," defines words in terms of their relationships to general concepts. The

Table III

Causes for Nonretrieval of Relevant Answers
by Narrow Semantic Code Programs

Number of Answers Missed

Ques. No.	Terminology	Role Indicators	Combinations	Levels	Abstracting Errors	Infixes	Program Errors	Obscurity	Other
A. Answers Retrieved by B-programs:									
1	3	2							
2	10								
3	38	12		5	2			1	
4		2							
5	1	2							
6	4	5	9	1			4	1	
7									
8	1		1						
9			4						
10	3			4		9			
11									
12	1	4	2	2	1				
13	7	8	15	11	2			1	
14									
Sub-total	68	35	31	23	5	9	4	3	
B. Answers Retrieved by C-programs:									
1	2	2		1					
2	3				2				
3	3	1			1				
4									
5									
6					1				
7	2	3		2		1			
8	3	2	2	2					
9	1		4		4				
10	1			1	5		1		
11									3
12									
13	2				1				
14									
Sub-total	17	8	6	6	14	1	1	0	3
TOTAL	85	43	37	29	19	10	5	3	3

Table IV

Causes for Retrieval of Nonrelevant Answers

by Narrow Semantic Code Program

Number of Answers Missed

Ques. No.	Information not in CA	Point of view	Role Indicators	Inflixes	False Combinations	Semantic Code	Add'l. Terminology	Minor Reference	Coding Error	Abstracting Error
1										
2	10	3			1		1	1	2	
3	8				2			1		
4	7	35								
5										
6										
7	9	2								
8										
9	5			30	4		2			
10	15	2			21	21				
11	3		32							
12	18	23								1
13										
14										
TOTAL	75	65	32	30	28	21	3	2	2	1

Sometimes concepts relating to a word are not included in its code and, as a result, this word is not linked theoreti-ally in the reverse code dictionary to other words incorporating this concept. The word is then not included in appropriate programs and relevant answers are missed.

Examples:

A. Question No. 1: High school dropout is related to juvenile delinquency.

Relevant section of T: .KAF dropout
 .KAL experience
 .KIS juvenile
 court

The words "juvenile court experience" were not included in the S-program as being related to "delinquency." No theoreti-cal relationship is indicated by the semantic codes:

delinquency DUD 200 2121
 juvenile FFL 1317
 court LAM 1102 1130

actual relationship of word to factor is made explicit by an alphabetical "infix" inserted between the first and second letter of each code. For example, the word "student" is coded "PAPL SWCH 2001." P-PL is the semantic factor signifying "human being"; S-CH is the semantic factor for "education" in general. The "A" infix means that a student is a human being (identity relationship); the "W" infix means that a student is acted upon by, or acts upon, education (affective relationship). A listing of the commonly used infixes is given in The Library of Tomorrow--Today.¹⁰ The number "2001" distinguishes the code for "student" from codes for other words which might combine the factors PAPL and SWCH (for example, "apprentice"). Some 200 semantic factors are used to encode words in the education research file. These are listed in The Library of Tomorrow--Today.¹¹

Sometimes concepts relating to a word are not included in its code and, as a result, this word is not linked thesaurally in the reverse code dictionary to other words incorporating this concept. The word is then not included in appropriate programs and relevant answers are missed.

Examples:

- A. Question No. 1: High school dropout as related to juvenile delinquency.

Relevant section of TA: ..KAP dropout
 missed by this program are: .KAL experience
 .KIS juvenile
 court

The words "juvenile court experience" were not included in the N-program as being related to "delinquency." No thesaural relationship is indicated by the semantic code:

delinquency	DVDD PASS 2131
juvenile	PYPL.1317
court	LACN LYGL 2130

"Delinquency" and "court" have no semantic factors in common; therefore, the reverse code dictionary shows no thesaural relationship. However, for this questioner, the terms are related, and the term for "delinquency" should have included some such semantic factor as "LXGL" (negation of Such legality), which would have associated it with "court."

B. Question No. 2: Pupil directed learning (independent study or discovery).

The relevant section of the telegraphic abstract has "individual study" rather than "independent study." The codes show no relationship:

Many of the independent RYPR.1098
retarded rather than individual SWPT.1101

C. Question No. 3: Influence of parents on their children's school performance.

The need for greater terminological control by the semantic code can be seen from the following list of codes assigned to the terms included in this program to represent the concept "school performance":

achievement	P-RD.1101
performance	P-SS.1006
behavior	P-SS.1171
success	P-RD.1103
accomplishment	P-RD.1344
learning	S-CH.1101
understanding	N-LL.1784
GPA	R-NG.1102

Terms not included in the N-program which occurred in relevant abstracts
The semantic codes for "deaf" and "hearing" do not exhibit any thesaural
missed by this program are:

intellectual	NYLL.1113
development	GART.1003
dropout	TYRM.1131
reporting	DACM.1055
ranking	RUNG.1697
school	LACN SUCH 2003
work	PASS.1105
adjustment	CANG.1026

Relevant section of telegraphic abstract: KAL audience practice

D. Question No. 5: Education of migratory children. *participation*

A relevant answer discusses the education of "Spanish speaking children of farm workers in Texas." Here it is the combination of terms, rather than any single term, which indicates the relationship to "migrant." Such an answer presents retrieval difficulties even with a stricter vocabulary control, since its relationship to the question depends on knowledge outside the abstract. *biology. Two people coding the same word in an educa-*

E. Question No. 6: Role of teaching machines in treating mental illness.

Many of the relevant answers to this question dealt with mentally retarded rather than mentally ill children. Again, the two terms have no semantic factor in common:

achievement	F-ED (product)
mentally ill	P-33 (DYSG MVMT 2169)
mentally retarded	SM (NWBT.1046 NVLL.1104 2100)
	M-IN (motion)

For this questioner the two terms are closely related, whereas the code stresses the difference between illness and retardation.

F. Question No. 8: Use of audiovisual aids in teaching deaf children.

The nature of the population was not specified in a relevant telegraphic abstract (KVV deaf or defective hearing), but indicated by the factors. Use of more multi-factor codes increases the possibility of false following phrase: ..KEC lipreading combinations on the 2-level, but these might be restricted, to some extent. The semantic codes for "deaf" and "hearing" do not exhibit any thesaural relationship with "lipreading":

stuttering	CUMT.1102 NVMT.1022 2184
deaf	BVPS.1002 DWSC 2025
hearing	NUTR.1100 SWND 2101
lipreading	CUMT.1318

G. Question No. 13: Effectiveness of learner participation in teaching importance in implementation these recommendations, when fully tested. motor skills by film.

Relevant section of telegraphic abstract: .KAL audience practice

(GURP.1026) and "practice" (PASS.1007).

Recommendations:

antic code, but occurs in any indexing system. However

C-NG (change)

G-RT (growth)

P-RD (product)

P-SS (process)

M-SM (amusement)

M-TN (motion)

To allow for the questioners' multiple points of view, it would seem

wise, wherever appropriate, to use the maximum number of semantic factors

(four) in each code.* Another possibility would be to revise the semantic

factor structure so that there would be less overlap in meaning among

factors. Use of more multi-factor codes increases the possibility of false

factors. Use of more multi-factor codes increases the possibility of false

combinations on the Z-level, but these might be restricted, to some ex

stuttering CUMT.1102 NWBT.1022 2184

CUMT.1102 NWBT.1022 2184

(speech) (handicap)

A. Question No. 3: Influence of parents on their children's school performance

not CUMT NWBT 2184

(communication) (stopping, retardation)

The S-program asks for (communication) (stopping, retard

*Updating (correction) programs have a top priority due to their

importance in implementing these recommendations, when fully tested.

2. Role Indicators

A role indicator is a code attached to each significant word indicating its function. The meanings of the most commonly used role indicators are listed below:

KEJ	population or thing acted upon and/or studied
KWV	attributes given
KIS	environmental or geographic location, by name or type
KEW	person interviewed or answering questionnaire
KAM	major process carried out on, by, or in relation to KEJ
KAH	condition of process
KQJ	media or agent of process
KEC	curriculum or material taught
KUP	attribute or behavior found
KAL	influencing factor
KWC	that towards which an attitude is noted
KXL	not an influencing factor
KWJ	product; device or material prepared
KAB	type of literature; design of study
KIT	year of original publication

Answers may be missed when the role indicators specified in the N-program are not the ones which appear in relevant telegraphic abstracts.

Examples:

A. Question No. 3: Influence of parents on their children's school performance.

The N-program asks for	..KAP performance (or related word)
	..KAL parent (or related word)
A relevant abstract has	..KAM predicting
	..KQJ parent
	attitude
	scale

F. Question No. 13: Effectiveness of learner participation in teaching

..KAP academic
motor skills achievement
.KAL attitude

The N-program asks for KAL participation.

B. Question No. 4: Decision-making.

The N-program requests the role indicators KUP or KEC with "decision."

A relevant telegraphic abstract has ..KAM training
..KQJ decision
making
assembly

Recommendations:

C. Question No. 5: Education of migratory children.

The N-program requests KEC or KAM.SUCH to indicate that education is discussed. A relevant abstract has ..KEJ migrant
..KEJ, KWJ educational
program
service

D. Question No. 6: Role of teaching machines in treating mental illness.

The N-program asks for KQJ teaching
machine

A relevant abstract has ..KUP KAP education
mentally retarded
teaching
machine

E. Question No. 7: Student interests and how they relate to achievement.

One frequently recurring role indicator combination in programs is

KAP - KAL (Property influenced - Influencing factor). For example, in

question 7, the program asks for ..KAP achievement
..KAL interest

On the other hand, answers judged relevant by the questioner had the role given in Table V. Twenty-six out of 103 relevant answers, or about 24%, were missed by specifying role indicators. On the other hand, by

omitting role indicators we increased the total number retrieved from 411 to 1098. With role indicators, ..KUP vocational
educational
interest

*Updating procedures (see note, page 18) must be applied to eliminate these inconsistencies.

F. Question No. 13: Effectiveness of learner participation in teaching motor skills by film.

The N-program asks for KAL participation.

A relevant abstract has ..KEC motor task
 .KQJ film
 .KAH participation

Recommendations:

If role indicators are to be effective as discriminatory devices in programs, the strictest control must be exercised in their application in the telegraphic abstract. One problem with the present file has been the frequent changes of policy in the use of role indicators--an unavoidable consequence of setting up a new system.* Thus, for example, in the beginning KAD was used for machine or device; now this role indicator is replaced by KQJ. KAH (condition) was used in the early file, eliminated, and then reintroduced. In early abstracts the teaching of a subject (A) by a particular method (B) was represented by ..KAM teaching
 .KQJ B
 .KEC A

In more recent abstracts the same information is represented by

..KEC A
 .KQJ B

If we consider only those questions for which the B-programs differ from the N-programs in the use of role indicators, we have the results given in Table V. Twenty-six out of 107 relevant answers, or about 24%, were missed by specifying role indicators. On the other hand, by omitting role indicators we increased the total number of answers retrieved from 411 to 1098. With role indicators, 11% of the retrieved answers were relevant; without role indicators 8% of the retrieved answers were relevant.

*Updating procedures (see note, page 18) must be applied to eliminate these inconsistencies.

Table V

Effect of Role Indicators in the Following Questions;
 N-Programs Were Broadened by Eliminating Role Indicators*

Ques. No.	No. of Relevant Answers			No. of Relevant Answers Eliminated by Role Indicators	Total No. of Answers	
	Total	N-Program	B-Program		N-Program	B-Program
1	6	0	4	2	1	16
4(a)	6	5	6	1	115	197
4(b)	13	11	13	2	115	197
5(a)	9	7	9	2	9	25
5(b)	2	1	2	1	9	25
6	12	0	1	1	0	67
8	12	3	3	2	0	6
12(a)	15	12	15	3	81	229
12(b)	11	7	11	4	81	229
13	21	0	19	8	0	89
14	0	0	0	0	0	18
TOTAL	107	46	85	26	411	1098

*In some cases B-programs employed other broadening devices besides the elimination of role indicators. Thus there may be a discrepancy between the number in column 5 and the difference between the numbers in columns 4 and 3.

Level

Level indicators, or punctuation, set up barriers within the telegraphic abstract between the various logical units of the discussion. Just as paragraphs and periods group thoughts in printed text, the position of the telegraphic abstracts corresponding to each level number is given below:

Level No.	Includes
1	semantic factor and end marker
2	coiled word (i.e., single space of telegraphic abstract)

3. Combinations

Most question programs can be regarded as conjunctions of semantic code and role indicator requirements at certain levels, although a few, question 2 for example, are simple disjunctions or strings of alternate terminology. Some relevant answers are missed because too many conjuncts are required; in other words, the program asks for A.B.C., but relevant answers contain only A.B or A.C or B.C. Usually this situation occurs when the original statement of the question was not sufficiently precise, and thus this problem is related to No. 8--obscurity.

Examples:

Question No. 6: Role of teaching machines in treating mental illness.
Question No. 11: High school dropout as related to juvenile delinquency.
Documents discussing the use of teaching machines in teaching the mentally ill were considered relevant, so that the original program requirement KAN.T-RT (treatment or therapy) should not have been specified.

Recommendations:

Many of these problems can be eliminated by a closer working relationship between questioner and programmer.

4. Levels

Level indicators, or punctuation, set up barriers within the telegraphic abstract between the various logical units of the discussion, just as paragraphs and periods group thoughts in printed text. The portion of the telegraphic abstract corresponding to each level number is given below:

Level No.	Includes
1	semantic factor and end number
2	coded word (i.e., single space of telegraphic abstract)

- 4 role indicator and coded word or words assigned to it
- 5 "phrase" of related words with their role indicators
- 6 "sentence" of related phrases
- 8 entire abstract

Role indicators commonly grouped together in phrases are listed in The Library of Tomorrow--Today.¹²

Answers may be missed because a combination of terms occurs in the telegraphic abstract at a level higher than that requested in the program. This problem is related to role indicators, since only certain role indicators are grouped together at a 5 level.

Example:

Question No. 1: High school dropout as related to juvenile delinquency.

The program asks for "dropout" or related terms and "delinquency" or related terms at the 5 level.

A relevant answer has

..KAM
..KQJ

investigation
school

etc. Several relevant answers surfaced, including the words "program" (PALC.1003) or "programming" (PULC.1003). alternate inflexes retrieved such as "program" or "programming" in the "curriculum" (PALC.1003 SYCH 2001).

..KAL
..KAP

characteristic
continuing
leaving
school

Recommendations:

It is difficult to separate the problems of terminology and role indicators from that of levels. It appears that the previously mentioned increased control over terminology and role indicators would go far in eliminating answers missed because of level barriers.

codes are specified.

5. Abstracting Errors

In a file of 4800 documents, some errors and omissions will inevitably occur. Two errors might be mentioned in particular:

(1) One phrase was set off from the body of the abstract by a 6-level ($\square\dots$), although it obviously related to the rest of the abstract.

(2) Significant words in the conventional abstract or title were omitted from the telegraphic abstract.

6. Infixes

A related term appears in the telegraphic abstract, but is coded with an infix different from that specified in the program.

Example:

Question No. 10: Programmed learning as compared with other teaching methods.

The program asks for "programmed" (FALC.1003) teaching, textbook, etc. Several relevant answers contained, instead, the words "program" (FALC.1003) or "programming" (FALC.1003). Unfortunately, allowing these alternate infixes retrieved much nonrelevant material dealing with "programs" of study, "programming" in the sense of scheduling, and "curriculum" (FALC.1003 SYCH 2001).

Recommendations:

Here again, more terminological control appears to be the answer. Different types of programs should be distinguished via the semantic code words used in the telegraphic abstract. More generally, care should be taken, in writing programs, that all infixes contained in relevant codes are specified.

7. Programming Errors

When the universal sign (-) is specified in the infix position of a semantic factor, for example N-LL, the infix V (NVLL) is not retrieved. Occasionally relevant terms using the V infix are overlooked, when the V infix is not specified separately in the program.

8. Obscurity

Since no contact was made with the questioner after the original statement of the question, inevitably some misunderstandings arose concerning the scope and meaning of the inquiry. In three cases, this led to a program which omitted relevant answers.

9. Other Causes

Three of the answers retrieved by the C-program should also have been retrieved by the semantic code programs. Since no reason could be found for nonretrieval, it was assumed that some human or machine error had occurred prior to receipt of the N and B program answers.

Factors Causing Retrieval of Nonrelevant Answers by N-Programs

We are concerned here with an analysis of the nonrelevant answers retrieved by the narrow program, since the generic nature of the broad program inevitably results in a high proportion of nonrelevant answers. If nonrelevant answers to the N-programs can be reduced to a minimum, it may be feasible to send out answers without preliminary evaluation.

In determining the factors causing retrieval of nonrelevant answers, both the system operation and the evaluation and comments of the questioner were taken into consideration. Nonrelevant answers resulting from system operation are recognizable and therefore potentially preventable through

system modification. Answers not recognized as nonrelevant from the original statement of the question, but so evaluated by the questioner, present a problem in question analysis rather than system operation. Both categories are considered below.

1. Omission of Information in the Conventional Abstract

Not all of the indexing terms in the telegraphic abstract appear in the conventional abstract which the questioner used for evaluation. As a result, it may not be clear to the questioner why certain answers were retrieved, and usually he will mark them nonrelevant rather than go to the trouble of checking an original source. Had he known the additional information in the telegraphic abstract, however, he might have considered the answer relevant. User tests should aim at this point.

Approximately one quarter of the nonrelevant answers to the N-program were rejected by the questioner for this reason. It is therefore recommended that the conventional abstract be expanded to include all information indexed in the telegraphic abstract. In this connection, while it may be noted that nine of the fourteen questioners commented on the difficulty in determining relevance from the information given in the conventional abstract, the balance between "indicative" and "informative" abstracts needs further investigation.

2. Point of View

The abstract relates to the content or at least the words of the question, but considers it from a point of view or in a context not of interest to the questioner.

Examples: material. In above, the use of two or more codes for the

A. Question 4: Decision-making.

It appears that the questioner was interested in the processes of decision-making and problem-solving from the point of view of game

theory rather than in practical situations such as "decision-making responsibility of college administrators" or "career decisions of high school students" or "problem-solving techniques in arithmetic."

B. Question No. 7: Student interests and how they relate to achievement.

Abstracts dealing with the "interest" of a film or TV program and its relationship to the student's learning from it were not relevant. In other words, interest must be a measured or stated attribute of the student rather than the media.

C. Question No. 12: Teaching of concepts by inductive or deductive techniques.

Experimental situations where an inductive vs. a deductive technique, broadly interpreted, is not the independent variable, were not of interest: for example, situations in which concept formation itself is the independent variable (educability as measured by concept formation) or situations comparing concept learning from film vs. filmstrip, etc.

Sometimes the problem seems to lie more in the ambiguity of words rather than in point of view. In this same question--No. 12--the writer of an article on "curriculum revision in the light of new mathematical concepts" is using the term "concept" in a looser sense than the questioner.

Recommendations:

It is difficult, in a word-based system, to make any general recommendation for elimination of answers which consider the subject from a point of view of no interest to the questioner. In some cases, greater precision in defining the question would have eliminated much of the nonrelevant material. In others, the use of two or more codes for the same word, where the word has more than one meaning, or elimination from

the telegraphic abstract of all occurrences of a word in a diffuse, as opposed to a specific, meaning would have been useful. Unfortunately, assignment of two codes to one word necessitates a further manual encoding procedure. An operational compromise should be sought.

3. Role Indicators

For some questions, the presently used role indicators are not sufficiently selective to distinguish relevant and nonrelevant answers.

Examples:

Question 11: Teaching of film (film appreciation, literary or comprehension).

Question 11 is the only one of the 14 where role indicators posed a serious problem in retrieving nonrelevant answers. These answers were concerned with teaching by film rather than teaching of film. In the program, a KEC or KEL role indicator was specified with "film;" and this latter one caused difficulty. KEL, in this case, means "thing studied," but no distinction is made between a study of the use of film in teaching (nonrelevant) and a study of the film itself (relevant). It is difficult to say, in this example, whether we might have been justified in restricting the role indicator to KEC, since no relevant answers were retrieved.

4. Infixes

Nonrelevant answers may be retrieved because a universal rather than a specific infix was used with some semantic factor. Usually this problem can be remedied in the program by a closer attention to the infixes of retrievable words in the reverse code dictionary whenever a universal infix is contemplated (but see No. 6 in the preceding section).

Example:

Question No. 4: Decision-making.

Two of the nonrelevant answers could be eliminated by specifying a V or F

The N-program contains the code N-LL.1108, which will retrieve the following terms:

decision	NWLL.1108	
indecision	NXLL.1108	
judgment	NALL.1108	
judging	NULL.1108	
judgmental	NYLL.1108	
undecided	NXLL.1108	
judged	NWLL.1108	
jury	NULL.1108	GARP 2100
adjudicator	NULL.1108	PAFL 2937

By specifying the infix W we eliminate such obviously nonrelevant words as "jury," "adjudicator," etc., as well as peripheral words such as "indecision." The net result of infix specification, in this case, was to eliminate one relevant answer and 18 nonrelevant answers. Use of infixes depends to some extent on user needs: i.e., completeness vs. relevancy.

5. False Combinations

Generally, the use of levels and role indicators reduces the probability of false combinations of terms. It is not surprising, then, that these proved a real problem only in Question 10, where no role indicators were used.

Example:

Question 10: Programmed learning as compared with other teaching methods.

Difficulty arose with the 4-level combination: R-FR.1228

S-CH

included in the program to retrieve such terms as "self instruction" on two levels. In 21 nonrelevant answers retrieval was caused by such 4-level combinations as:

teacher	student	student teacher
self	self	self
judgment	evaluation	control

Ten of the nonrelevant answers could be eliminated by specifying a Y or U

infix in S-CH; all of them by using end numbers to restrict the combination to the specific codes for "self" and "instruction." This latter change, however, would have eliminated three relevant answers.

5. Semantic Codes

The semantic codes assigned to a group of terms are not sufficiently selective to make distinctions of interest to the questioner.

Example:

Question 10: Programmed learning as compared with other teaching methods.

The 2-level combination "NACH.SUCH" is used in codes for programmed teaching devices, self-scoring devices and training devices such as link trainers or gunnery training devices. The questioner did not consider these last two categories to be appropriate and designated as nonrelevant abstracts about the following devices:

- MSU serial reactor
- Toronto CC
- Task Trainer
- Voice Visualizer
- Link Trainer
- Hushburn Apparatus
- Double Disk Pursuit Apparatus
- Humboldt
- SRA Reading Rate Accelerator
- Trainer Tester
- Purdue Speech Intensity Demonstrator
- all gunnery training devices.

These answers cannot be eliminated in reprogramming, unless we list the code for each relevant "NACH.SUCH" device separately--an impractical procedure because of the number of codes involved. Any solution involves recoding--including the code "FWLC.1003" in all codes for programmed teaching machines, for example.

By employing some of the discriminatory features such as role

7. Additional Terminology

For most questions, the programmer adds words to the program which are not in the original statement of the question but appear to be related to it. In three cases--"honors program" in question 2 and "reliability" and "validity" in question 9--these terms proved nonrelevant to the questioner.

8. Minor Reference to Subject

In two cases documents were apparently rejected because the subject matter of the question occupied a relatively insignificant place in the total article. If all information indexed in the telegraphic abstract appeared in the conventional abstract seen by the evaluator, it is of course possible that a greater number of articles would have been rejected for this reason.

9. Coding Errors

Two errors in the automatic encoding process occurred in question 2, where the code for "honors program" was assigned to the word "honors." Such errors result from incomplete matching of dictionary and word cards and are rare.

10. Abstraction Errors

Only one occurred--in question 12, where a nonrelevant answer was retrieved because "self-concept" was written on two lines instead of one.

Conclusions

The alternate search strategies, in general, showed the semantic code-telegraphic abstract retrieval system capable of a wide range of possible relevancy-recall ratios. By broadening semantic code programs--i.e., by employing fewer of the discriminatory features such as rule

indicators, levels and infixes--the recall factor is increased. By applying more of the discriminatory features the relevancy factor is increased. The relevancy and recall factors of the faceted classification lie somewhere between those for a very broad and a very narrow semantic code program.

The faceted classification schedule proved useful in retrieving answers to questions stated in relatively specific terms, where the semantic code did not bring out all possible thesaural relationships. For this reason, these classifications were applied to all documents in the file.

The sample of 14 questions proved, in most cases, too small to indicate a significant difference at the .05 level between relevancy and recall factors of broad and narrow semantic code programs and between these and faceted classification programs. Also, it is now felt that the classification into broad and narrow semantic code programs is not sufficiently discriminating. More useful comparisons will be obtained if a series of programs is written, each one differing from the others for that question by one feature only. Then the relevancy and recall potential of each discriminatory feature can be determined.

Many difficulties in satisfying the users would have been eliminated by a closer contact with the questioner. The programmer worked with the original statement of the question only, and the results obtained, in many cases, reflected as much a lack of communication between questioner and systems operator as a malfunctioning of the system.

For all these reasons, it is recommended that the investigations be extended on a more discriminatory basis to a larger set of questions and

to a user group with whom there is a maximum of personal contact and cooperation.

The examination of nonretrieved relevant answers and retrieved non-relevant answers, in relation to the narrow semantic code programs, led to the following recommendations for further investigation:

1. Modification of the semantic code to achieve more consistent terminological control.
2. Strict control should be exercised in the application of role indicators and levels in the telegraphic abstract.
3. A procedure for automatically updating the file should be developed so that when changes in the use of role indicators, levels or codes are made they may be applied to material already in the file.
4. The extent to which information indexed in the telegraphic abstract should be included in the conventional abstract must be carefully tested and considered.

PART II

RELEVANCY STUDY

In most studies of retrieval systems the term "relevancy" is taken as an undefined concept or given a definition such as "a relevant document is one with direct bearing on the question" or "a relevant document is one which satisfies the questioner." In these cases, relevancy is being determined subjectively by the questioner or other evaluator.

It is also possible to define relevance objectively: i.e., a relevant document is a member of a set defined by certain objective and possibly quantifiable characteristics. For example, in the comparison of retrieval systems by Cleveland¹³ and Fels¹⁴, the relevant document is the source document for the question; in other words, trial questions submitted to the system are phrased from documents known to be in the system. Another type of objective and, in this case quantifiable, relevance is the "relevance number" of Maron and Kuhns.¹⁵ In their "probabilistic indexing" the index terms characterizing the subject content of the document are weighted numerically. Given an information request, these weightings are used to compute a relevance number which is a measure of the probability that the document will satisfy the questioner. Objective criteria of relevance are usually based on a special situation, such as questions with predetermined answers or an intimate and detailed knowledge of user needs, as in a research library serving a small clientele. Since neither of these conditions will be true of an operational Educational Research Information Center, it seemed preferable, in this study, to obtain subjective assignments of relevance by the questioner and then to attempt to determine the objective properties of these relevant documents.

In determining these objective criteria two questions were asked:

- (1) If answers are evaluated by two different evaluators, how much will assignments of relevance vary?
- (2) Do the properties which characterize relevant documents lie in the language of the answer, as opposed to such extra-linguistic factors as the number of answers scanned, the order in which answers are scanned, or the form of the original article? If linguistic properties can be isolated, to what extent are they reflected in our indexing-searching apparatus?

The investigations resulting from each question will be discussed separately.

Variations in Relevancy

Our purpose was to determine the variation in evaluation of answers by two evaluators: first, when one is a CDCR staff member and the other the original questioner; and second, when one is the original questioner and the other a second outside evaluator interested in the same question.

Procedure

A total of 1214 abstracts, representing answers to 14 questions, were evaluated by the original questioners and returned to the Center. Four of the 14, representing 320 answers, were given two outside evaluations. All answers sent to questioners had been evaluated previously at the Center as relevant or peripheral (answers judged nonrelevant were not sent). Evaluation categories were defined as follows:

- relevant: an article which has direct bearing on the question
- peripheral: an article of possible or limited use in relation to the question
- nonrelevant: an article which has no bearing on the question

It should be understood that many of the answers termed "peripheral" in the internal CDCR evaluation were not answers which would ordinarily be sent to a questioner, but simply answers with some connection, no matter how remote, to the question. In view of this latitude, it is unlikely that more than a small percentage of the answers not sent to the questioner would have been of interest.

One might also make the following distinction between relevant, peripheral and nonrelevant: relevant answers are those we prefer to send to the questioner; peripheral answers are those about which we are indifferent; nonrelevant answers are those we prefer not to send to the questioner.

In most practical searching operations, the system operator is faced with the problem of whether or not to send answers which do not fit the questioner's specifications exactly but are related to them (peripheral answers). There are two possibilities for making a wrong decision: first, if he sends peripheral answers which the questioner considers nonrelevant; and second, if he does not send peripheral answers the questioner might have considered relevant. It was therefore determined, with the sample, whether the number of wrong decisions would have been greater if answers considered peripheral by the CDCR evaluator had not been sent.

Results

Results for all 14 questions are given in Table VI; results for the four questions given two outside evaluations in Table VII. The changes in evaluation from the CDCR evaluation to the original questioner evaluation are given in Table VIII; differences in evaluation between the two outside evaluators in Table IX; differences in the three evaluations in Table X.

Table VI
Results for All 14 Questions
Total Answers: 1214

Evaluated By	Relevant		Peripheral		Nonrelevant	
	No.	%	No.	%	No.	%
COCE	703	58	311	42	--	--
Questioner	474	39	285	24	455	37
Other	129	11	125	10	130	11
Original Questioner	47	4	27	2	196	16
Second Questioner	427	35	258	22	259	21

Table VII
Results for 4 Questions
Given 2 Outside Evaluations

Total Answers: 320

<u>Evaluated by</u>	<u>Relevant</u>		<u>Peripheral</u>		<u>Nonrelevant</u>	
	No.	%	No.	%	No.	%
CDCR	132	41	188	59	--	--
Original questioner	45	14	85	27	190	59
Second questioner	85	27	59	18	176	55

Table VIII
 Changes in Evaluation, All 14 Questions

Evaluated by		No. of Answers	% of Answers
Evaluation by		No. of Answers	% of Answers
CDOR	Questioner		
Relevant	Relevant	401	33
Relevant	Peripheral	139	12
Relevant	Nonrelevant	163	13
Peripheral	Relevant	73	6
Peripheral	Peripheral	146	12
Peripheral	Nonrelevant	292	24
Nonrelevant	Nonrelevant	218	17

Table XX

Changes in Evaluation, Outside Evaluators

<u>Evaluated By</u>		<u>No. of Answers</u>	<u>% of Answers</u>
<u>Original</u> <u>Questioning</u>	<u>Second</u> <u>Questioning</u>		
Relevant	Relevant	18	6
Relevant	Peripheral	29	11
Peripheral	Relevant		
Relevant	Nonrelevant	19	18
Nonrelevant	Relevant		
Peripheral	Peripheral	19	6
Peripheral	Nonrelevant	71	23
Nonrelevant	Peripheral		
Nonrelevant	Nonrelevant	118	37

Table X

Changes in Evaluation, 3 Evaluators

	<u>Evaluated By</u>	<u>No. of Answers</u>	<u>% of Answers</u>
<u>CDCR</u>	<u>Outside</u> <u>Evaluators</u>		
Relevant	Relevant	15	3
Peripheral	Peripheral	6	3
Relevant	Nonrelevant	26	8
Peripheral	Relevant by one evaluator	37	12
Peripheral	Nonrelevant	89	28

(1) An answer not which both questioners considered irrelevant.

(2) An answer not which at least one questioner considers relevant. By sending both relevant and peripheral answers a wrong decision was made in 33% of the answers. 22 were rated relevant by the COCR and nonrelevant by both questioners and 11% were rated peripheral by the COCR and nonrelevant by both questioners. Had the peripheral answers not been sent, a wrong decision would have been made with 70% of the answers. 22 were rated relevant by the COCR and nonrelevant by both questioners, and 11% were rated peripheral by the COCR and relevant by at least one questioner. Applying the Wilcoxon matched-pairs signed rank test to the difference scores for the sample of 14 questions, there is no difference at the .05 level in percent wrong decisions made by sending peripheral answers, percent wrong decisions made by not sending peripheral answers.

CONCLUSIONS

That more answers were considered relevant by the COCR evaluator than by the original questioner is not surprising, since it may be assumed in

By sending both relevant and peripheral answers, as evaluated by the CDR, a wrong decision was made with 37% of the answers: 12% of the answers considered relevant in the CDR evaluation and 24% of the answers considered peripheral were called nonrelevant by the questioner. On the other hand, if only the answers considered relevant by the CDR evaluator had been sent, a wrong decision would have been made 19% of the time. As before, 12% of the relevant answers were judged nonrelevant and, in addition, 6% of the peripheral answers were considered relevant.

Considering only the four questions given two outside evaluations, results are similar. A wrong decision is defined as follows:

- (1) an answer sent which both questioners consider nonrelevant
- (2) an answer not sent which at least one questioner considers relevant.

By sending both relevant and peripheral answers a wrong decision was made in 36% of the answers: 8% were rated relevant by the CDR and nonrelevant by both questioners and 28% were rated peripheral by the CDR and nonrelevant by both questioners. Had the peripheral answers not been sent, a wrong decision would have been made with 20% of the answers: 8% were rated relevant by the CDR and nonrelevant by both questioners, and 12% were rated peripheral by the CDR and relevant by at least one questioner. Applying the Wilcoxon matched-pairs signed ranks test to the difference scores for the sample of 14 questions, there is no difference at the .05 level in percent wrong decisions made by sending peripheral answers vs. percent wrong decisions made by not sending peripheral answers.

Conclusions

That more answers were considered relevant by the CDR evaluator than by the original questioner is not surprising, since it may be assumed in

this situation where there was little contact between occasional system operator and questioner that the question is defined more precisely by the latter. Some of the differences in evaluation would certainly be eliminated by greater questioner-operator interaction.

For an information center serving a selected clientele, continuing contact between questioner and system operator is usually not economically feasible, so that some decisions about relevancy will be necessary. The above results indicate that fewer wrong decisions are made when peripheral answers are not sent (19% as compared to 37%), but these figures must be interpreted in terms of the relative cost to the customer of missing a relevant answer and receiving a nonrelevant answer. As was pointed out in Part I, these costs may vary and not be equal.

The evaluations by two external evaluators would indicate that there is disagreement about the relevance of answers among questioners themselves, a disagreement amounting in this case to 19% of the answers (i.e., answers evaluated relevant by one questioner, nonrelevant by the other). In three of these four questions, the two questioners had no contact with each other; with the fourth, both were investigators on the same project. In this last case, none of the answers were evaluated relevant by one evaluator and nonrelevant by the other, and 19% of the answers were given the same evaluation (as opposed to a total of 49% for all four questions).

In general, these results indicate that one cannot say an answer is relevant or nonrelevant in any absolute sense. It is more useful to assign it a probability of relevance, and in the next section an attempt is made to find some quantitative measure of this property.

Characteristic of Relevant Answers

The purpose in attempting to isolate some property or properties of relevant, as opposed to peripheral and nonrelevant, answers is to simplify and expedite the identification of these answers within the present retrieval system. If relevancy is based on criteria external to the indexing and retrieval system--as total number of answers scanned or order in which answers are scanned--general rules for identification of these answers will be difficult to formulate. Certainly, from this sample, there does not appear to be a correlation between number of answers sent for a question and proportion of answers judged relevant, as the following figures will indicate (here, as before, (a) and (b) represent two outside evaluations of the same question):

Question #	No. of Answers	% of Answers Judged Relevant	
		CDCR Evaluation	Questioner's Evaluation
08	3	0	100
01	6	83	83
14	7	29	0
06	8	63	100
05(a)	12	83	67
05(b)	12	83	17
07	21	62	32
09	36	50	19
13	39	69	49
11	53	6	0
02	61	61	41
04(a)	84	33	7
04(b)	84	33	15
12(a)	101	40	15
12(b)	101	40	11
03(a)	123	44	13
03(b)	123	44	48
10	333	98	80

The rank-difference correlation between size of output and percent of answers evaluated relevant by the CDCR is only $-.03$; between size of output and answers evaluated relevant by the questioner, $-.37$.

As a first hypothesis, for purposes of this study, it was postulated that some linguistic property--some characteristic of the language of an answer--distinguishes the relevant answers. More specifically, the hypothesis is tested that it is the relative frequency of the question words and words related to the question words in the various forms the answer may take--title, abstract, index to document--which characterizes relevant answers.

Procedures

As described in Part I, the education pilot research file consists of a tape file of telegraphic abstracts; i.e., listings of the significant words of an article coded in a semantic code and organized into a meaningful structure by levels and role indicators. In order to ask a question of the tape, a program is written which requires various configurations of words, role indicators and connectives. The words in the program include one or more of the following types:

- (1) significant words of the question.
- (2) words related thesaurally to the question words by the semantic code (i.e., including at least one of the same semantic factors as the related question word).
- (3) words not related to question words by the semantic code, but known by the programmer, through intuition or knowledge of the file, to be related to the question words.

We shall call these three types respectively question words, semantic code references, and program additions.

To clarify this categorization, the words of several of the programs are classified below according to type:

Question 1: High school dropout as related to juvenile delinquency.

Question Words:	Word	Code
	High school	LACH*SUCH*2101
	Dropout	TYRM.1131
	Juvenile	PTPL.1317
	Delinquency	SVIC*PASS*2131

Semantic Code References:		
	Junior high school	LACH*SUCH*2102
	Secondary school	LACH*SUCH*2103
	Dropping out	TYRM.1131
	Termination	TARM.1001
	Terminating	TYRM.1001
	Terminal	TYRM.1001
	Discontinuance	TARM.1010
	Boundary	"
	Crime	QASD*PASS*2393

Program Additions:		
	Withdrawal	RARE.1002
	Withdrawn	TYRM.1002
	Continuance	CANT.1010

Question 8: Influence of parents on their children's school performance.

Question Words:		
	Influence	CANT.1100
	Parent	RARE.1104
	Children	PTPL.1004
	School	LACH*SUCH*2003
	Performance	PASS.1006

Semantic Code References:		
	Parental	RYEL.1104
	Mother	RARE.1003
	Maternal	RYEL.1003
	Father	RARE.1103
	Paternal	RYEL.1103
	Performing	PASS.1006
	Performed	PASS.1006
	Performer	PASS.1006*PAFL
	Behavior	PASS.1171
	Behaving	PASS.1171
	Behavioral	PASS.1171
	Behavioristic	"
	Conduct	PASS.1171
	(all types of educational institutions)	LACH*SUCH
	Pupil	PAFL*SMCH*2001
	Student	"
	Learner	"

Program Additions:

Home	LACN.1105
Achievement	PARD.1101
Achieving	PURD.1101
Achieved	PWRD.1101
Underachievement	PXRD.1101
Overachievement	PYRD.1101*RAPR*2564
Achiever	PYRD.1101*PAPL*2100
Underachiever	PXRD.1101*PAPL*2100
Nonachiever	"
Overachiever	PVRD.1101*RYPR*PAPL*3564
Score	PWRD.1101*LAMN*2100
Scoring	PWRD.1101*LUMN*2100
Tallying	PWRD.1101*LUMN*2262
Scored	PWRD.1101*LUMN*(L))
Machine scored	PWRD.1101*LUMN*MQCH*3249
(names of achievement tests)	PYRD.1101*DACH*MUSR
Success	PARD.1103
Successful	PYRD.1103
Unsuccessful	PXRD.1103
Accomplishment	PARD.1344
Failure	PARD.1348
Failed	PWRD.1348
Learning	SUCH.1101
Learned	SWCH.1101
Overlearning	SUCH.1101*RYPR*2565
Relearning	SUCH.1101*RYPR*2798
Understanding	NULL.1784
Comprehending	"
Comprehension	NALL.1784
Misunderstanding	NXLL.1784
Understandability	NYLL.1784*GAPR*2313
Comprehensibility	"
Understandable	NYLL.1784*GYPR*2313
Grade	RANG.1102
Graded	RWNG.1102
Grading	RUNG.1102
Nongraded	RXNG.1102
Nongrading	RXNG.1102*PUSS*2382
Point Hour Ratio	RYNG.1102*RWPR*LAMN*3213
Grade Point Average	RYNG.1102*RWPR.1214*LAMN*3133
Senior	RYNG.1102*RYPR*2100
Junior	RYNG.1102*RYPR*2712
Kindergarten	RANG.1102*RYPR*2136

The frequency of these three types of word--(1) question words, (2) semantic code references, (3) program additions--in relevant, peripheral and nonrelevant answers was computed and used to determine the

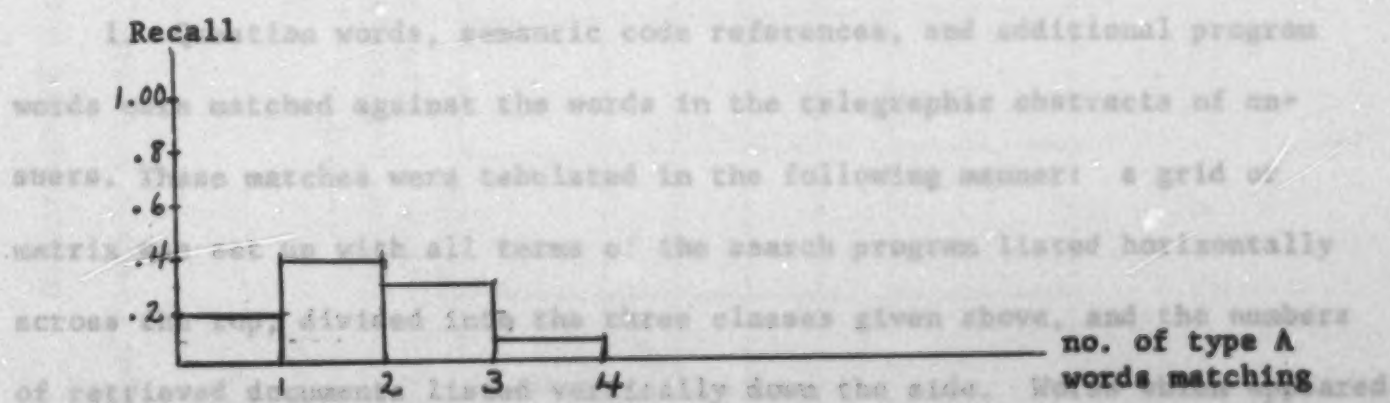
effectiveness of each in

- (a) locating relevant answers (i.e., as indexing terms)
- (b) indicating the relevance of answers (i.e., as evaluating terms).

The effectiveness of a word "A" as an indexing term is rated by the ratio of number of relevant answers indexed by "A" to total number of relevant answers. This ratio corresponds to the recall factor considered in Part I. One may extend the concept of indexing effectiveness to types of words as well as single words; the types of words being question words, semantic code references, and program additions. With word types we must consider not just a single ratio but a frequency function plotting number of type "A" words against the observed probability of this number of type "A" words occurring in relevant answers; i.e.,

$$\text{Probability of X words} = \frac{\text{no. of relevant answers with X words}}{\text{total relevant answers}}$$

Thus one may indicate the indexing effectiveness of type A words by a histogram such as the following:



The distribution of words of type B may be plotted and compared with words of type A. If the mean number of words of type B is greater than the mean for type A and if this difference is significant, then words of type B articles, prepositions, relative pronouns, etc.--even where the words

are said to be more effective as indexing terms than words of type A.

The relevancy factor of Part I--i.e.,

$$\frac{\text{no. of relevant answers containing term A}}{\text{no. of answers containing term A}}$$

indicates the evaluating effectiveness of a single term "A" since it answers the following question: if a questioner receives all answers containing word "A", what proportion of the answers will be relevant? This concept may also be extended to types of words by a function showing the evaluating effectiveness of 1, 2, 3 and 4 words of type "A". This function will not be a frequency function, since the probabilities do not total one. However, if the evaluating effectiveness graph for words of type A is everywhere above that for type B, we may conclude that words of type A are more effective in evaluating than words of type B.

Similarly, this study compares the relative effectiveness of title, conventional abstracts and telegraphic abstracts as indexing and evaluating tools by comparing the frequency of question words in each of the three.

The procedure, in detail, was as follows:

1. Question words, semantic code references, and additional program words were matched against the words in the telegraphic abstracts of answers. These matches were tabulated in the following manner: a grid or matrix was set up with all terms of the search program listed horizontally across the top, divided into the three classes given above, and the numbers of retrieved documents listed vertically down the side. Words which appeared in the telegraphic abstract for each document were then simply checked off. As an example, Table XI shows the matrix for question 1. Words included in the tabulation included all words of the question except function words--articles, prepositions, relative pronouns, etc.--even where the words

Table XI

Occurrence of Question Words, Semantic Code References,

and Program Additions in Telegraphic Abstract Answers to Question 1

Question Words			Semantic Code References								Program Additions				
Document Numbers	High School	Drop-out	Delinquency	Secondary School	Junior High	Drop-out	Termination	Terminal	Discontinuation	Boundary	Crime	Delinquent	Withdrawal	Withdrawn	Continuance
Relevant Answers															
S0123			✓												
S0433	✓	✓	✓												
S1968		✓													
T0152			✓									✓			
T0374															
Nonrelevant Answers															
E0167	✓	✓													

seemed quite indefinite in meaning. For example, in question 10-- programmed learning as compared to other teaching methods--the question words included programmed, learning, compared, teaching, methods.

2. A count was made of the number of words with 0, 1, 2, 3, and 4 or more (a) question words, (b) semantic code references, and (c) program additions in relevant, peripheral and nonrelevant answers, as determined by the questioner.

3. Counts were similarly made of the number of conventional abstracts and of titles containing 0, 1, 2, 3, and 4 or more question words. For this count it was found most useful to interpret "question word" as the actual form in the question or any grammatical variant of it: e.g., delinquency or delinquent; dropout, dropping out, or dropped out.

4. The following ratios, or probabilities, were then computed for $X = 0, 1, 2, 3, 4$ or more:

$$P(R) = \frac{\text{total relevant answers}}{\text{total answers}}$$

$$P(XQA) = \frac{\text{no. of answers with } X \text{ or more question words in telegraphic abstract}}{\text{total answers}}$$

$$P(XSA) = \frac{\text{no. of answers with } X \text{ or more semantic code references in TA}}{\text{total answers}}$$

$$P(XPA) = \frac{\text{no. of answers with } X \text{ or more program additions in TA}}{\text{total answers}}$$

$$P(XTA) = \frac{\text{no. of answers with } X \text{ or more words matches of any kind in TA}}{\text{total answers}}$$

Indexing Effectiveness Ratios

$$P(XQA/R) = \frac{\text{no. of relevant answers with } X \text{ or more question words in TA}}{\text{total relevant answers}}$$

$$P(XQA/Ph) = \frac{\text{no. of peripheral answers with } X \text{ or more question words in TA}}{\text{total peripheral answers}}$$

$$P(XQA/N) = \frac{\text{no. of nonrelevant answers with X or more question words in TA}}{\text{total nonrelevant answers}}$$

and similarly

$$\begin{array}{lll} P(XSA/R) & P(XSA/Ph) & P(XSA/N) \\ P(XPA/R) & P(XPA/Ph) & P(XPA/N) \\ P(XTA/R) & P(XTA/Ph) & P(XTA/N) \end{array}$$

Evaluation Effectiveness Ratios

$$P(R/XQA) = \frac{\text{no. of relevant answers with X or more question words in TA}}{\text{total answers with X or more question words in TA}}$$

$$= \frac{P(R) P(XQA/R)}{P(XQA)} \quad \text{by Bayes formula}$$

and similarly for semantic code references, program additions and total words

$$\begin{array}{l} P(R/XSA) \\ P(R/XPA) \\ P(R/XTA) \end{array}$$

The probabilities for X question word matches in the title and conventional abstract are similarly computed and the indexing and evaluating effectiveness probabilities designated as $P(XQTi/R)$ and $P(R/XQTi)$ for titles and $P(XQC/R)$ and $P(R/XQC)$ for conventional abstracts.

5. $P(XQA/R)$ -- the probability of X question word matches given a relevant answer--indicates the indexing effectiveness of question words, and in the same way $P(XSA/R)$ and $P(XPA/R)$ indicate the indexing effectiveness of semantic code references and program additions. By plotting these values for $X = 0, 1, 2, 3$, and 4 or more a function was obtained indicating the indexing effectiveness of each type. Similarly, by plotting the values of $P(R/XQA)$, $P(R/XSA)$ and $P(R/XPA)$ a function was obtained indicating the evaluating effectiveness of each type of word. Functions representing the indexing and evaluating effectiveness of titles and conventional abstracts were similarly obtained.

Table XII

6. The significance of the following differences was tested by Chi square:

- (a) differences in distribution of question words, semantic code references and program additions in

(1) relevant vs. peripheral answers

(ii) peripheral vs. nonrelevant answers

- (b) difference in distribution of question words in conventional abstract, title and telegraphic abstract in

(1) relevant vs. peripheral answers

(ii) peripheral vs. nonrelevant answers

No. of Words	All Words			Question Words			Semantic Code Refs.			Program Additions		
	Relevant	Peripheral	Nonrelevant	Relevant	Peripheral	Nonrelevant	Relevant	Peripheral	Nonrelevant	Relevant	Peripheral	Nonrelevant
1 or more	.87	.60	.62	.46	.43	.33	.04	.03	.04	.71	.45	.39
2 or more	.13	.40	.38	.54	.57	.67	.96	.97	.96	.29	.55	.61
3 or more	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
4 or more	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Mean	2.92	2.16	2.02	1.46	1.57	1.03	.29	.34	.31	1.19	.61	.58
St. Dev.												

Results: A. Question Words, Semantic Code References and Program Additions

The columns for relevant answers in Table XII show the relative effectiveness of question words, semantic code references and program additions as indexing terms for our sample of 1214 answers, made up of 474 relevant, 285 peripheral and 455 nonrelevant answers. Table XIII shows the relative effectiveness of question words, semantic code references and program additions as evaluating terms.

The mean number of matches for each type indicates that question words are most effective as indexing terms, followed by program additions. Semantic code references appear to have comparatively little indexing value. Table XIII shows the surprising result that up to combinations of two or more, the program additions are more effective than question words in evaluating answers. One suspects this anomaly results from the comparatively small number of answers with program addition matches as compared with the number having question word matches (613 vs. 904) and might disappear with a larger sample.

Table XII

Comparison of Question Words, Semantic Code References
and Program Additions as Indexing Terms
and Program Additions as Evaluating Terms

No. of Words	AU Words			Question Words			Semantic Code Refs.			Program Additions		
	Rele- vant	Peri- pheral	Nonre- levant	Rel.	Per.	Nonrel.	Rel.	Per.	Nonrel.	Rel.	Per.	Nonrel.
1 or more	.86	.71	.66	.25	.30	.47	.71	.45	.39			
2 or more	.87	.60	.62	.46	.43	.35	.04	.05	.04	.39	.12	.12
3 or more	.60	.32	.26	.14	.09	.04	0	.01	0	.08	.03	.05
4 or more	.31	.15	.11	.02	.02	0	0	0	0	.01	.01	.02
Mean	2.92	2.16	2.02	1.48	1.25	1.05	.29	.36	.51	1.19	.61	.58
St. Dev.	1.30	1.37	1.10	.94	.89	.90	.53	.61	.57	.96	.81	.89

Table XIII

Comparison of Question Words, Semantic Code References
and Program Additions as Evaluating Terms

No. of Words	Probability of Relevance (Relevance Factor)			
	All Words	Question Words	Semantic Code References	Program Additions
1 or more	.39	.45	.29	.52
2 or more	.47	.44	.39	.66
3 or more	.57	.61	---	.58
4 or more	.57	.78	---	.42

At the .01 level, the frequency of total words, question words and program additions was found to be greater in relevant answers than in peripheral or nonrelevant answers; the frequency of total words, question words and program additions did not differ significantly between peripheral and nonrelevant answers. The frequency of semantic code references does not differ between relevant and peripheral answers, but is significantly greater in nonrelevant than in peripheral or in relevant answers.

It appears that relevant answers differ the most from peripheral and nonrelevant answers in total number of word matches, rather than by a particular type of word match. To compare these differences more exactly, a distance function is defined to measure the distance between relevant and peripheral and peripheral and nonrelevant answers in terms of total word matches, question word matches, etc. The distance between relevant and peripheral answers in terms of question word matches (d_{qp}) is defined as

$$d_{qp} = \sqrt{\sum_{i=1}^n [P(QWA/R) - P(QWA/P)]^2}$$

where the values of the indexing probabilities are taken from Table XIII.

One may similarly define the distance between relevant and peripheral

Table XII also indicates that the probability of occurrence of words of all types is greater for relevant answers than for peripheral or nonrelevant answers. Peripheral answers are similarly distinguished from nonrelevant answers, but the differences are not so great. The distribution of each type of word or of total words in relevant, peripheral and nonrelevant answers can be compared by applying a Chi square analysis to each pair of distributions. Such an analysis determines whether relevant answers are characterized by a particular type of language.

At the .01 level, the frequency of total words, question words and program additions was found to be greater in relevant answers than in peripheral or nonrelevant answers; the frequency of total words, question words and program additions did not differ significantly between peripheral and nonrelevant answers. The frequency of semantic code references does not differ between relevant and peripheral answers, but is significantly greater in nonrelevant than in peripheral or in relevant answers.

It appears that relevant answers differ the most from peripheral and nonrelevant answers in total number of word matches, rather than by a particular type of word match. To compare these differences more exactly, a distance function is defined to measure the distance between relevant and peripheral and peripheral and nonrelevant answers in terms of total word matches, question word matches, etc. The distance between relevant and peripheral answers in terms of question word matches (d_{QRPh}) is defined as

$$d_{QRPh} = \sqrt{\sum_{x=0}^4 [P(XQA/R) - P(XQA/Ph)]^2}$$

where the values of the indexing probabilities are taken from Table XIII.

One may similarly define the distances between relevant and peripheral

answers in terms of semantic code references and program additions and total words, and the same distances between relevant and nonrelevant answers. The following results support the conclusion that relevant answers are most strongly characterized by total word matches:

Distance between relevant and peripheral answers in terms of		
No. of Words Required by Program	Type of Output Which Will Be Retrieved	% of Output Which Will Be Relevant
	question words	.16
	semantic code references	.05
1 or more	program additions	.38
	total	.42

Distance between relevant and nonrelevant answers in terms of		
No. of Words Required by Program	Type of Output Which Will Be Retrieved	% of Output Which Will Be Relevant
	question words	.25
	semantic code references	.22
1 or more	program additions	.42
	total words	.52

Taking Tables XII and XIII in conjunction, one can predict, on the basis of this sample, what results will be obtained with various program types, as shown in Table XIV.

Some general conclusions may be drawn from the data of Table XIV:

1. Question words are most useful in locating relevant answers but cannot be employed exclusively if completeness is desired, since 14% of the relevant answers contained no question word at all.
2. Semantic code references are least effective in programs either in retrieving relevant answers or in distinguishing relevant from peripheral and nonrelevant answers.
3. The greatest drop in percent of relevant answers retrieved, considering words of all types, is between combinations of two and three words.

Table XIV

**Results of Various Types
of Program and Search Output**

No. of Words Required by Program	Type of Words Required by Program	% of Relevant Answers Retrieved	% of Output Which Will Be Relevant
1 or more	Any type	100	39
	QW	86	43
	SCR	25	29
	PA	71	52
2 or more	Any type	87	47
	QW	46	44
	SCR	04	39
	PA	39	66
3 or more	Any type	60	57
	QW	14	61
	SCR	0	--
	PA	08	58
4 or more	Any type	31	57
	QW	02	78
	SCR	0	--
	PA	01	42

From the means of the relevant answer columns in Table XV, it is apparent that words in the conventional abstract are the best indexes to relevant answers. However, differences between the two types of abstracts in relevant answers are small compared with the difference between these and titles, the title words having a much lower indexing effectiveness. On the other hand, title words are most effective on evaluating terms--an expected result since these generally are the most significant of the

If at least two words are required, 87% of the relevant answers will be retrieved; if at least three words are required, 60% of the relevant answers will be retrieved.

4. The greatest drop in percent of relevant answers retrieved by question words occurs between combinations of one and two words. If at least one word is required 86% of the relevant answers will be retrieved; if two or more words are required 46% of relevant answers will be retrieved.

5. The program additions appear to be the most effective evaluating terms at least in combinations of one and two, although as has been suggested this rather surprising result may be a peculiarity of this sample. Further tests along these lines should be carried out.

Results: B. Titles, Conventional Abstracts and Telegraphic Abstracts

In comparing words in titles, conventional abstracts and telegraphic abstracts, it was considered best to admit any grammatical variant of a question word (e.g., teaching, teacher, taught, teachable), since in consulting a conventional index it is likely that a questioner would "look up" under all forms. The relative effectiveness of titles, conventional abstracts and telegraphic abstracts as indexes is shown in Table XV; as evaluating tools in Table XVI.

From the means of the relevant answer columns in Table XV, it is apparent that words in the conventional abstract are the best indexes to relevant answers. However, differences between the two types of abstracts in relevant answers are small compared with the difference between these and titles, the title words having a much lower indexing effectiveness. On the other hand, title words are most effective as evaluating terms--an expected result since these generally are the most significant of the

Table XV

Comparison of Title Words, Conventional Abstract Words,
and Telegraphic Abstract Words as Indexing Terms

Number of Question Words	In Title			In Conventional Abstract			In Telegraphic Abstract		
	Rel.	Per.	Nonrel.	Rel.	Per.	Nonrel.	Rel.	Per.	Nonrel.
1 or more	.65	.39	.37	.88	.63	.61	.90	.80	.81
2 or more	.22	.11	.07	.60	.27	.30	.54	.49	.45
3 or more	.05	.02	0	.25	.07	.08	.19	.15	.13
4 or more	0	0	0	.04	0	0	.02	.03	.02
Mean	.92	.54	.45	1.77	.97	.91	1.65	1.47	1.41
St. Dev.	.84	.73	.89	1.04	.92	1.04	.94	1.03	.98

Table XVI

Comparison of Title Words, Conventional Abstract Words
and Telegraphic Abstract Words as Evaluating Terms

Number of Question Words	Probability of Relevance		
	In Title	In CA	In TA
1 or more	.52	.48	.42
2 or more	.61	.57	.44
3 or more	.65	.65	.49
4 or more	---	---	---

A Chi-square test of the difference in distribution of words in relevant vs. peripheral and peripheral vs. nonrelevant at .01 level shows that for all three indexes--titles, telegraphic abstracts and conventional abstracts--the probability of question words is greater in relevant than in peripheral or nonrelevant answers but does not differ significantly in peripheral and nonrelevant answers. Thus it may be concluded that the frequency of question words in titles, conventional abstracts and telegraphic abstracts distinguishes relevant from nonrelevant answers. The distance function defined earlier may again be used to indicate the relative strength of these differences:

Distance between relevant and peripheral answers in terms of:

titles	.28
conventional abstracts	.45
telegraphic abstracts	.11

Distance between relevant and nonrelevant answers in terms of:

titles	.32
conventional abstracts	.44
telegraphic abstracts	.14

indexing terms. The lower evaluating effectiveness of the telegraphic abstract is partly due to the fact, as noted in Part I, that the questioners do not see the telegraphic abstracts, only titles and conventional abstracts.

This same situation may explain the high frequency of question words in the telegraphic abstracts of peripheral and nonrelevant answers. If an appropriate term appears in the conventional abstract the evaluator will see it and call the answer relevant; if the same term appears in the telegraphic abstract only, he may well call the same answer nonrelevant.

A Chi-square test of the difference in distribution of words in relevant vs. peripheral and peripheral vs. nonrelevant at .01 level shows that for all three indexes--titles, telegraphic abstracts and conventional abstracts--the probability of question words is greater in relevant than in peripheral or nonrelevant answers but does not differ significantly in peripheral and nonrelevant answers. Thus it may be concluded that the frequency of question words in titles, conventional abstracts and telegraphic abstracts distinguishes relevant from nonrelevant answers. The distance function defined earlier may again be used to indicate the relative strength of these differences:

Distance between relevant and peripheral answers in terms of:

1. In proper titles	.28
conventional abstracts	.45
telegraphic abstracts	.12

Distance between relevant and nonrelevant answers in terms of:

2. The titles	.32
conventional abstracts	.44
telegraphic abstracts	.14

Based on these results, the following conclusions are drawn:

1. Titles alone have a low indexing effectiveness. If only titles are searched, at most 65% of the relevant answers will be retrieved, whereas if telegraphic abstracts are searched, 90% of the relevant answers will be retrieved.

2. Indexing terms which appear in the telegraphic abstract but not in the conventional are of little value.

Conclusions

The main interest in the investigations described in this section lies more in the method than in the actual results. This particular sample might be criticized on two grounds:

(1) It is too small.

(2) It is biased, since a heavy proportion of the total answers come from one question: Question 10, 333 out of 1214 answers.

Unfortunately, the delays inherent in sending answers outside the CDCR for evaluation made it inconvenient to increase the sample size when it became apparent that all of the original 24 questioners were not returning answers.

However, the following conclusions might be drawn, even from this imperfect sample:

1. In programming questions, the number of words required in combination is of greater significance than the type of words required as far as retrieval of relevant as opposed to peripheral or nonrelevant answers is concerned.

2. The related words which the programmer adds to the original statement of the question are useful as indexing and evaluating words,

although their indexing effectiveness is lower than that of question words.

3. The semantic code leads from the question words to very few additional indexing terms, and those which it selects are more likely to retrieve peripheral and nonrelevant than relevant answers. Theoretically, the semantic code should have related the question words not only to the semantic code references but also to the program additions. Since it did not, it should be investigated whether an improved semantic code which leads to a greater number of thesaural relationships than the present one will have a higher indexing and evaluating value. This conclusion reinforces that reached in the studies described in Part I.

4. Because of their low indexing effectiveness, titles need not be recorded on tape as separate entities.

5. All of the indexing information in the telegraphic abstracts should be included in the conventional abstract.

In general, it is suggested that the method of quantitative assessment of indexing and evaluating effectiveness of terminology should be utilized in further studies, so that results here indicated may be substantiated and extended.

12. H. S. Maron, "Evaluation of the Performance of a Terminology Retrieval System by Modified Moore's Test," *American Documentalist*, vol. 14, no. 1 (January 1963), 28-34.

13. H. S. Maron and J. L. Kuhse, "On Relevance, Probability, Indexing and Information Retrieval," *Journal of the Association for Documentation*, vol. 1, no. 3 (1960), 215-244.

References:

1. Pamela W. Reeves, et al., The Library of Tomorrow--Today (Cleveland: Center for Documentation and Communication Research, 1962), pp. 46-52.
2. Joseph Becker and Robert M. Hayes, Information Storage and Retrieval: Tools, Elements, Theories (New York: Wiley, 1963), p. 251.
3. Reeves, ch. 4.
4. Maurice F. Tauber and Oliver L. Lilley, Feasibility Study Regarding the Establishment of an Educational Media Research Information Service, (New York: Columbia University, 1960), pp. 82-184.
5. Reeves, pp. 289-295.
6. Cyril Cleverdon, "The Cleverdon-WRU Experiment: Conclusions," in Center for Documentation and Communication Research, Information Retrieval in Action (Cleveland: Western Reserve University Press, 1963), p. 102.
7. Cleverdon, pp. 106-107.
8. Becker and Hayes, p. 256.
9. Sidney Siegel, Nonparametric Statistics for the Behavioral Sciences (New York: McGraw-Hill, 1956), pp. 75-83.
10. Reeves, pp. 232-233
11. Reeves, pp. 239-245.
12. Reeves, p. 171.
13. Cleverdon, pp. 101-107.
14. Eberhard M. Fels, "Evaluation of the Performance of an Information Retrieval System by Modified Mooers Plan," American Documentation, vol. 14, no. 1 (January 1963), 28-34.
15. M. E. Maron and J. L. Kuhns, "On Relevance, Probabilistic Indexing and Information Retrieval," Journal of the Association for Computing Machinery, vol. 7, no. 3 (1960), 216-244.

Appendix A: An Example of Alternate Strategies

Broad Semantic Program (B-Program)

Question No. 1: High school dropout as related to juvenile delinquency.

Faceted Semantic Code Programs (F-Program)

<u>Letter Symbol</u>	<u>Code</u>	<u>Words Retrieved</u>
A	LACN*SUCH*2101	high school senior high senior high school
B	LACN*SUCH*2102	junior high junior high school
C	LACN*SUCH*2103	secondary school
D	KAL	
E	T-RM.1131	dropout dropping out
F	T-RM.1101	termination terminal terminating
G	T-RM.1010	discontinuance boundary
H	R-ML.1002	withdrawal withdrawn removed
I	KAP	
J	D-DD*P-SS*2131	delinquency delinquent nondelinquent
K	D-DD*P-SS*2393	crime

Narrow Program (N-Program)

$$^6\{(A+B+C) \cdot ^5[{}^4[D \cdot (E+F+G+H)]^4 \cdot {}^4[I \cdot (J+K)]^4]^5\}^6$$

Broad Program (B-Program)

$$^6\{[E+F+G+H] \cdot [J+K]\}^6$$

Broad Revised Program (BR-Program)

$$^6 \{ J + K \} ^6$$

Faceted Classification Program (C-Program)

<u>Letter Symbol</u>	<u>Code</u>	<u>Facet</u>
A	FLSJ	junior high school
B	FLSH	senior high school
C	FEDE	delinquency
D	APDO	dropout, continuance

$$^6 \{ (A + B) \cdot (C + D) \} ^6$$